SUPPLEMENTARY INFORMATION:

Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children

Kyle M. Walsh, Anand P. Chokkalingam, Ling-I Hsu, Catherine Metayer, Adam J. de Smith, Daniel I. Jacobs, Gary V. Dahl, Mignon L. Loh, Ivan Smirnov, Karen Bartley, Xiaomei Ma, John K. Wiencke, Lisa F. Barcellos, Joseph L. Wiemels, Patricia A. Buffler

Supplementary Methods

Ethics Statement:

This study was reviewed and approved by institutional review committees at the University of California Berkeley, the California Department of Public Health (CDPH) and all collaborating institutions. Written informed consent was obtained from all parent respondents.

Study Population:

The study includes Hispanic participants from the California Childhood Leukemia Study (CCLS) (1). CCLS includes children from 35 Northern and Central California counties, recruited between 1995 and 2008. Subjects with newly diagnosed leukemia were recruited from nine area hospitals within ~72 hours of diagnosis. Comparison with the California Cancer Registry (1997-2003) shows that the CCLS captured ~95% of children diagnosed with leukemia at the participating hospitals. Cases ascertained through the CCLS participating hospitals represented

~76% of all diagnosed cases in the study region. Birth certificate information obtained from the Office of Vital Records at the California Department of Public Health was used to select 1-2 controls for each case, matching on date of birth, sex, Hispanic ethnicity, and maternal race. The eligibility criteria for cases and controls were: 1) at least one parent self-reported Hispanic ethnic status 2) residency in the study area 3) no previous cancer diagnosis 4) less than 15 years of age at time of diagnosis (cases) or index date (controls).

Determination of immuno-phenotype and cytogenetic profiles:

Immuno-phenotype was determined for ALL cases using flow cytometry profiles. Those positive for CD19 or CD10 (≥20%) were classified as B-lineage and those expressing CD2, CD3, CD4, CD5, CD7, or CD8 (≥20%) were classified as T-lineage, as previously described (2). Ploidy was determined using G-banding or FISH, and *TEL-AML1* translocations were identified by fusion of the *TEL* and *AML1* loci, as previously described (2).

Genotyping:

Samples were genotyped at the UC Berkeley School of Public Health Genetic Epidemiology and Genomics Laboratory using the Illumina Human OmniExpress v1 platform, containing 730,525 SNP markers. DNA was isolated from dried bloodspots collected at birth and archived by the Genetic Diseases Screening Program of the CDPH. Newborn blood specimens were available for 87% of interviewed participants. Extraction was performed using the QIAamp DNA Mini Kit (QIAGEN, USA, Valencia, CA). Genotype reproducibility was verified using ten duplicate samples with average concordance across all genotypes > 99.99%. Genotype quality was high and SNP associations did not appear to suffer from biases related to population

stratification, as indicated by the low genomic inflation factor from the genomewide SNP association analyses (λ=1.025). Samples were screened for cryptic relatedness using 10,000 unlinked SNPs, and were excluded if the proportion of the genome identical-by-descent was > 0.15.

Calculation of individuals' ancestral components:

From the full set of SNPs genotyped in the Hispanic cases and controls, a linkage-reduced set of 63,303 autosomal SNPs, evenly distributed across the genome, was extracted. These SNPs include 129 ancestry-informative markers (AIMs) selected to distinguish African, European, and Native American populations as previously described and empirically tested (3). This SNP set was generated using a window size of 50 SNPs, shifting the window in 5-SNP steps, with a Variance-Inflation-Threshold of 1.43 (corresponding to R^2=0.30) in Plink v1.07 (4). This same set was extracted from the Human Genome Diversity Project (HGDP) genotype data, previously generated using an Illumina 650K array (5). All included SNPs passed aforementioned quality-control filters.

The genetic structure of study subjects was evaluated using the program Structure v2.3.1 (6). Percent membership in an ancestral group can distinguish membership in one group from membership in all others, and therefore may be superior to computing eigenvectors in situations where 3 or more ancestral populations contribute to a population's genomic structure, as is the case for Hispanics (7). The 372 unrelated HGDP individuals included 111 Africans (Bantu, Yoruba, San, Biako, Mandenka, Mbuti), 107 Native Americans (Pima, Maya, Karitana, Surui, Colombian), and 154 Europeans (Caucas Mountain Russians, Basque, Orkney, Sardinian, French, Italian) (5).

Because each ancestral proportion is a continuous variable ranging from zero to one, the odds ratio can be calculated for any increment of ancestry. We present odds ratios for each 20% increase in Native American ancestry as an easily interpretable level and a level approximately equal to the standard deviation of the Native American ancestry variable in our sample.

Calculating contribution of known susceptibility loci to ethnic incidence rate ratios:

Incidence rate ratios (IRRs) were calculated according to varying genotypic relative risks (GRR) and ethnic group allele frequencies using previously described methods (8). As an example, we start with the supposition that the GRR for ALL for persons with one risk allele is 2.00, and the GRR for persons with two risk alleles is 3.00 (relative to those homozygous for the "wild-type" allele). Assuming Hardy-Weinberg equilibrium at the population-level, if the frequency of the risk allele in Native Americans is 0.80 (p=0.80), then the proportions of AA (p^2), AB (2pq), and BB (q^2) genotypes are 0.64, 0.32, and 0.04, respectively. To calculate a normalized Native American incidence rate, the genotype proportion is multiplied by the associated GRR: 0.64 (3.00)+0.32 (2.00)+0.04 (1.00)=2.60. Given a European risk allele frequency of 0.20, the European normalized incidence rate is 0.04 (3.00)+0.32 (2.00)+0.64 (1.00)=1.40. In this example, the Native American/European IRR is 1.86 (2.60/1.40).

To perform these calculations, additive odds ratios for the previously identified risk SNPs were obtained from our association analyses of Hispanic B-cell ALL and substituted for the GRR. Risk allele frequencies for Native Americans, Europeans, and Africans were obtained from HGDP data, while risk allele frequencies for Hispanics were obtained from HapMap data(9). Confidence intervals were calculated using a permutation-based method which obtained odds ratios and allele frequencies from a Gaussian distribution. The distribution of population-

level allele frequencies were defined by the parameters (mean and standard deviation) observed in the HapMap/HGDP data, while the distribution of odds ratios was defined by the effect size parameters observed in the association analyses. From these distributions, 100,000 permutations were drawn to yield 100,000 IRR estimates. The fifth and ninety-fifth percentile of these IRR estimates were taken to represent the 95% confidence interval around the IRR calculated using study data.

Supplementary Results

Supplementary Tables:

Table S1: Demographic and tumor characteristics of Hispanic acute lymphoblastic leukemia cases and controls from the California Childhood Leukemia Study appearing in the association analyses.

	Case	Control	Total
Sample Size	297	454	751
n (%)			
Male	156 (52.5)	240 (52.9)	396 (52.7)
B-cell	297 (100.0)	NA	NA
Hyperdiploid (51+ chromosomes)	97 (32.7)	NA	NA
Tel-AML1	40 (13.5)	NA	NA
Neither Hyperdiploid nor *Tel-AML1*	118 (39.7)	NA	NA
Cytogenetic data not available	42 (14.1)	NA	NA
Mean (SD)			
Age	5.2 (3.3)	5.3 (3.4)	5.3 (3.4)
% Native American Ancestry	38.3 (15.7)	36.0 (16.3)	36.9 (16.1)
% African Ancestry	6.7 (5.7)	6.3 (5.0)	6.5 (5.3)
% European Ancestry	55.0 (16.2)	57.7 (17.2)	56.6 (16.8)

Table S2: Odds ratios (95% CI) for associations of genome-wide Native American ancestry, leukemia risk loci identified in previous GWAS and case-control status among Hispanic CCLS participants.

Covariate	OR (95% CI)[1]	P[1]	OR (95% CI)[2]	P[2]	OR (95% CI)[3]	P[3]	OR (95% CI)[4]	P[4]	OR (95% CI)[5]	P[5]	OR (95% CI)[6]	P[6]
%Native American ancestry	1.20 (1.00-1.45)	**0.048**	1.22 (1.01-1.47)	**0.036**	1.17 (0.97-1.41)	0.091	1.15 (0.91-1.46)	0.082	1.12 (0.92-1.35)	0.27	1.19 (0.99-1.43)	0.070
rs4132601-G (*IKZF1*)	1.46 (1.16-1.83)	**$1.3x10^{-3}$**	-	-	1.45 (1.15-1.82)	**$1.7x10^{-3}$**	1.46 (1.16-1.84)	**$1.2x10^{-3}$**	1.51 (1.19-1.92)	**$8.0x10^{-4}$**	1.46 (1.16-1.84)	**$1.4x10^{-3}$**
rs3731217-T (*CDKN2A*)	1.76 (1.17-2.65)	**$4.6 x10^{-3}$**	1.74 (1.16-2.62)	**$5.9x10^{-3}$**	-	-	1.76 (1.17-2.64)	**$5.0x10^{-3}$**	1.85 (1.21-2.83)	**$3.1x10^{-3}$**	1.79 (1.19-2.71)	**$4.0x10^{-3}$**
rs7088318-A (*PIP4K2A*)	1.19 (0.94-1.51)	0.14	1.21 (0.95-1.53)	0.12	1.18 (0.93-1.50)	0.16	-	-	1.19 (0.93-1.52)	0.16	1.19 (0.94-1.51)	0.15
rs7089424-C (*ARID5B*)	2.33 (1.85-2.92)	**$2.6x10^{-14}$**	2.36 (1.88-2.97)	**$1.5x10^{-14}$**	2.35 (1.87-2.96)	**$1.8x10^{-14}$**	2.32 (1.85-2.91)	**$3.2x10^{-14}$**	-	-	2.34 (1.86-2.94)	**$2.2x10^{-14}$**
rs2239633-G (CEBPE)	1.35 (1.09-1.68)	**$6.6x10^{-3}$**	1.34 (1.08-1.67)	**$8.1x10^{-3}$**	1.34 (1.08-1.67)	**$8.5x10^{-3}$**	1.34 (1.07-1.66)	**$8.8x10^{-3}$**	1.37 (1.09-1.72)	**$6.3x10^{-3}$**	-	-

[1] OR and P calculated in a logistic regression model adjusted for: age, sex and %African ancestry. Odds ratios are for each additional copy of the risk allele (for SNPs) or for each 20% increase in %Native American ancestry. [2] Additionally adjusted for genotype at rs4132601 (*IKZF1*). [3] Additionally adjusted for genotype at rs3731217 (*CDKN2A*). [4] Additionally adjusted for genotype at rs7088318 (*PIP4K2A*). [5] Additionally adjusted for genotype at rs7089424 (*ARID5B*). [6] Additionally adjusted for genotype at rs2239633 (*CEBPE*).

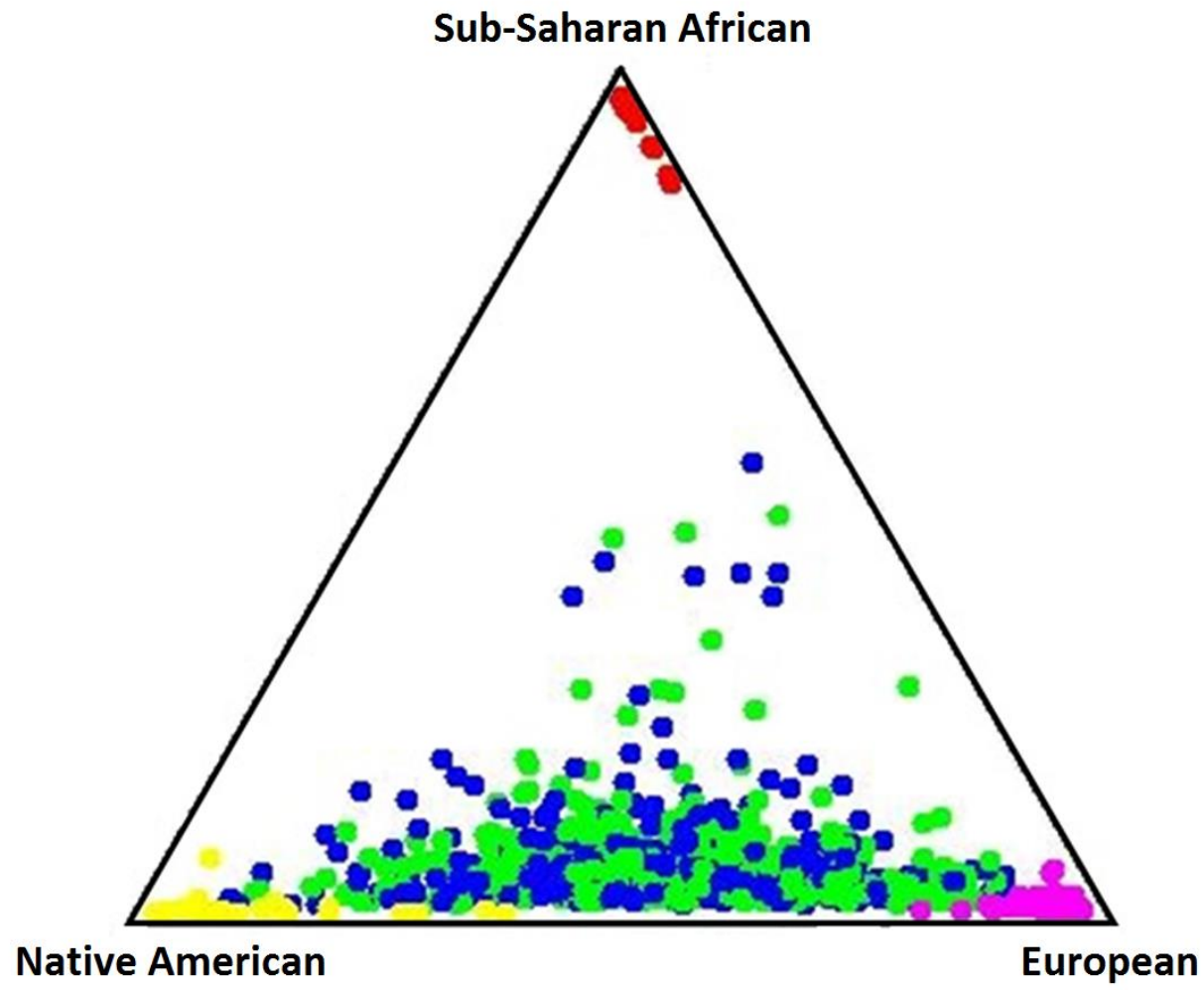
Table S3: SNP effect size, risk allele frequency and contribution to hyperdiploid B-cell ALL ethnic incidence rate ratios (IRR) by established susceptibility loci.

Risk Allele[1]	SNP Effect Size (95% CI)[2]	Risk allele frequency[3]				Hispanic-Caucasian IRR (95% CI)	Hispanic-African IRR (95% CI)
		Caucasian	African	Hispanic	Native American		
rs4132601-G (*IKZF1*)	1.50 (1.07-2.12)	0.301	0.212	0.160	0.224	0.892 (0.817-0.970)	0.957 (0.897-0.992)
rs3731217-T (*CDKN2A*)	2.46 (1.16-5.20)	0.863	0.907	0.880	1.000	1.014 (0.976-1.033)	0.978 (0.942-0.994)
rs7088318-A (*PIP4K2A*)	1.37 (0.96-1.96)	0.588	0.243	0.727	0.939	1.065 (0.997-1.132)	1.274 (0.991-1.577)
rs7089424-C (*ARID5B*)	2.91 (2.05-4.12)	0.304	0.235	0.380	0.570	1.139 (1.006-1.162)	1.303 (1.179-1.433)
rs2239633-G (*CEBPE*)	2.07 (1.44-2.98)	0.522	0.832	0.610	0.612	1.091 (1.005-1. 122)	0.826 (0.765-0.872)

[1] SNPs have previously been reported to increase risk for B-cell ALL in a published genome-wide association study. [2] Odds ratios are derived from the CCLS Hispanic case-control study, comparing 97 high-hyperdiploid B-cell ALL cases to 454 controls. [3] Allele frequencies for Caucasians, Africans and Native Americans are from Human Genome Diversity Panel data. Allele frequencies for Hispanics are from HapMap data.

Supplementary Figures:

Figure S1: Triangle plot showing estimated admixture in the Hispanic case-control sample.



Estimates were performed in Structure using 63,303 unlinked autosomal SNPs, including 129 AIMs, and data from the Human Genome Diversity Project on 111 African (Red dots), 154 European (Pink dots) and 107 Native American (Yellow dots) founders. The figure depicts ancestry in 297 Hispanic B-cell ALL cases (Blue dots) and 454 Hispanic controls (Green dots) from the California Childhood Leukemia Study.

Supplementary References:

1. Ma X, Buffler PA, Layefsky M, Does MB, Reynolds P. Control selection strategies in case-control studies of childhood diseases. *Am J Epidemiol* 2004; 159(10): 915-21.

2. Aldrich MC, Zhang L, Wiemels JL, Ma X, Loh ML, Metayer C *et al.* Cytogenetics of Hispanic and White children with acute lymphoblastic leukemia in California. *Cancer Epidemiol Biomarkers Prev* 2006; 15(3): 578-81.

3. Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M *et al.* Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet* 2012; 8(3): e1002554.

4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81(3): 559-75.

5. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; 319(5866): 1100-4.

6. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003; 164(4): 1567-87.

7. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006; 2(12): e190.

8. Jacobs DI, Walsh KM, Wrensch M, Wiencke J, Jenkins R, Houlston RS *et al.* Leveraging ethnic group incidence variation to investigate genetic susceptibility to glioma: a novel candidate SNP approach. *Front Genet* 2012; 3: 203.

9. The International HapMap Project. *Nature* 2003; 426(6968): 789-96.